FITTING MODELS TO DATA: THE MATHEMATISING STEP IN THE MODELLING PROCESS

Lídia Serrano,* Marianna Bosch,* Josep Gascón**

*Universitat Ramon Llull (Spain), **Universitat Autònoma de Barcelona (Spain)

This paper presents a mathematical modelling activity experienced with students of first year university level centred on a problem of forecasting sales using one-variable functions. It then focuses on the back and forth movements between the initial system – a time-series of the term sales of a firm – and the different models proposed to make the forecasting. The analysis of these movements, that are at the core of the 'mathematising step' of the modelling cycle, shows how the initial empirical system is being enlarged and progressively enriched with new variables and mathematical objects. Thus the development of a modelling activity initiated with a real-situation may soon lead to a process where the mathematising affects both the system and the model.

1. THE MATHEMATISING STEP IN THE MODELLING PROCESS

In current didactic contracts, the validity of the mathematical knowledge students have to learn usually has its last guarantee in an external source of the activity: the teacher. It is the teacher who, as a last resort, decides if a result is correct or wrong, if the used tool or technique was the best possible choice, etc. Because of this dominant epistemology underlying current didactic contracts of our teaching institutions, research in mathematics education puts forward an 'experimental epistemology' more in accordance with the Galilean's spirit of modern science. According to this epistemology, scientific knowledge (and mathematics in particular) is building up in permanent contrast with 'empirical facts' that, added to the principles of theoretical coherence, represent the main elements of proof. The reproduction of this 'experimental epistemology' in mathematics underlies the Theory of Didactic Situations (Brousseau, 1997), especially through the notion of adidactic situation and the principle of knowledge construction in contrast with a *milieu*. The recent developments of the Anthropological Theory of the Didactic (Chevallard 2004 and 2006) have introduced the notion of 'media and milieu dialectics' as an analysis tool of the necessary interaction between a *milieu*, i.e. any system devoid of any didactic intention, and the *media* (in the sense of 'mass media') as any source of information or pre-existent knowledge. The aim of this paper is to

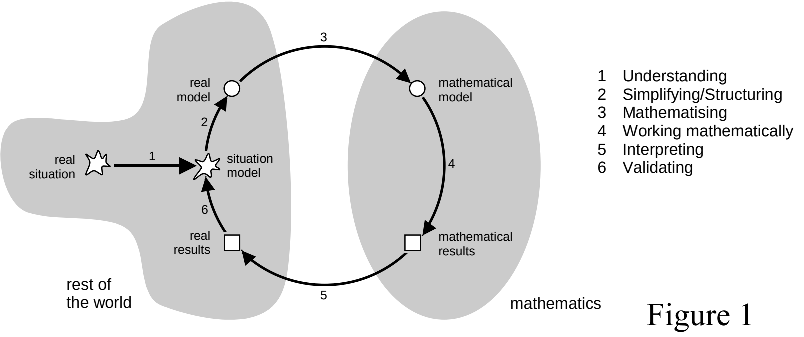


Figure 1

consider how these notions can help analyse a concrete step of the modelling process as it is considered in many research works in the ‘modelling and applications’ domain using the modelling cycle (Blum & Leiβ 2006), namely the ‘mathematising’ step (see figure 1).

This paper considers a special modelling activity that has been experimented with first-year students of a mathematics course for economy and business at university level. The real situation that is modelled is a problem of forecasting sales given the historical data or previous term sales. The concrete ‘mathematising’ of this situation consists in choosing an appropriate mathematical model (a one-variable function) fitting the empirical given data. The possibility of choosing different models and the need for a criterion to select one starts a process of contrast between the models and the empirical system acting as a ‘milieu’. The next section presents the conditions of the teaching experience and outlines the work of the students when approaching the sales forecast problem. The analysis of the experience in terms of the ‘media and milieu dialectics’ is detailed in the third section, before concluding about the importance of considering the ‘mathematisation’ of a mathematical system – that is, ‘intra-mathematical modelling’ – as a step of the modelling process analogue to those included in the modelling cycle.

2. A MODELLING WORKSHOP ON ‘FORECASTING SALES’

2.1. Conditions of the experience

The didactical experimentation we present here corresponds to a first course of mathematics in Economics Studies during the academic year 2006/07. It is important to underline that the teaching conditions of this course do not correspond to a traditional one. First, the university we refer to is a private university that organizes teaching in not very large groups (between 30 and 60 students) where every student has a personal laptop computer. Second, the course has been designed by a researcher in mathematics education and the experimentation was carried out by four teachers, three of whom are also researchers in didactics.

The course was designed drawing special attention to modelling activities. Its main goal, as it explicitly appears in the syllabus, is ‘to get students learn to elaborate and use mathematical models for the description, analysis and resolution of problematic situations that can be found in business, economy, finance or daily life. […] Students should be able to analyze problematic situation in terms of dependence between variables, pointing out the relevant information needed to construct a mathematical model of this situation. And they should know how to use the mathematical model proposed and how to synthesize the results obtained with these models in order to generate new knowledge and new questions about problematic situations considered.’

The programme is divided into three blocks that correspond to the three term periods of an academic year: linear algebra, calculus in one variable, and calculus in several variables/optimization. The course is structured in two weekly sessions of two hours:

the first one is a lecture (teachers' explanations and problem resolutions on the blackboard) and the second one is used to carry out a 'mathematical modelling workshop', centred on the study of a problematic question connected to the field of economy or business. The work here presented corresponds to the workshop experimented during the second term, within the domain of 'one variable calculus', which lasted 5 sessions.

The work at the workshop was organised in the following way: The students work in groups of 3 or 4 and have to write and present a weekly report about the partial results obtained at each session. At the end of the term, an individual final report has to be presented at the moment of the evaluation (a written exam which includes two different problems and a question related to the workshop). This exam represents 50% of the qualification; the written reports 40%, and the remaining 10% corresponds to the individual resolution of problems during some of the lectures.

2.2. The question of 'forecasting sales': analysis of its generative power

The initial question of the workshop was formulated as follows:

> A firm registers the term sales of its 7 main products during 3 years. They ask us the following questions:
>
> → What amount of sales can be forecasted for the next terms? Can we get a formula to estimate the forecasts? Which are its limitations and guarantees? How to explain them?
>
> → What products sales are increasing more than 10% a term? Less than 12% a term?

The data were 'prepared' by the teachers so that they correspond to seven elementary functions of different types (quadratic, cubic, rational, exponential) with an error term added.[1] The values of each function were slightly changed with the aim of distorting them, but without losing the general "tendency" of the original function.

The workshop's aim was to give students a problem close to a real situation where functions appear as a suitable model. Both the use of Excel in the first term of the course and the students' familiarity with elementary functions (it was the theme of the sessions just preceding the workshop) allowed them to initially detect a tendency in the sales (for example from a graphic representation of the data) and look for a function that fitted this tendency. The firm question proposed also included the idea of percent variation, which we expected would make the study of function variations appear during the workshop. Given that the workshop was run in parallel with the lectures on function derivatives, it was also expected that, at any time, the study of the sales' variation could be connected with them.

[1] The concrete functions were: $0{,}5(x - 6)^3 + 2000$; $2{,}5(x + 5)^2 + 100$; $5500/(x + 4)$; $1300 \cdot 085^x$; $1500 - 1200/(x + 1)$; $2{,}5(x + 5)^2 + 100$; $1300 \cdot 085^x$). The second experimentation in 2007/08 was carried out with 'real' data taken from some macroeconomic magnitudes of different countries: population, oil production, traffic crashes, unemployment rate, etc. The main difference between the two workshops appears in the study of the variations, because the real data have stronger fluctuations and do not always present a clear tendency.

The election of a sales forecast situation was mainly motivated by the fact that it enables to clearly distinguish between the economic system (sales) and the models used (functions). Moreover, working with different products needs to consider different models, raising the problem of the fitting between the model and the modelled system. In other words, the aim of the workshop was to make students use functions as a model of a simple economic system and quickly raise the question of the election of the model and its validation.

2.3. General organisation of the modelling workshop

We here report the four workshops experienced, corresponding to four classes of a (the) first-year course of mathematics for economics and business led by four different teachers working in team. Each group has a teacher, the same one for the lectures and the workshop sessions. All classes were prepared by the team and all sessions were discussed personally or by mail before and after being carried out. Each teacher, at the end of each workshop session, wrote a report in which he/she explained the development of the session, and sent it by mail to the other teachers.

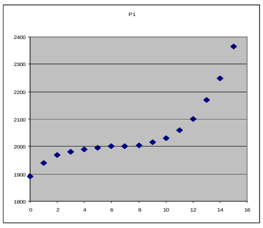
Before the workshop started, the students had four lectures dedicated to introducing the elementary families of functions, from straight lines to exponential functions. The students learned how to use the general expression of every family of functions and to associate them with different graphics. In other words, the students were taught how to assign an algebraic expression to the graphic of a function, among a set of given families. They saw how to deduce the graphic of $y = af(x - b) + c$, from the 'basic function' $y = f(x)$ and, reciprocally, how to deduce the expression of any function $y = af(x - b) + c$ given its graphic and knowing the original 'basic function' $y = f(x)$. The lectures given in parallel with the workshop introduced the notion of absolute and relative variation of a function between two points, the notion of the derivative's function, the notion of straight line tangent, etc. within the general problem of the study of variations. The functions considered were always related to economical situations, such as the *incomes* depending on the *sales*, the *cost* depending on the *production*, the *demand* depending on the *price*, etc.

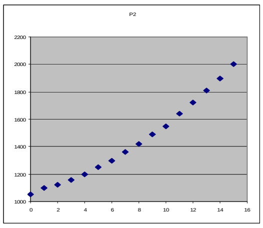
2.4. Description of the workshop sessions

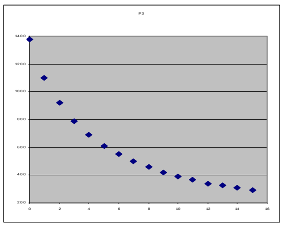
We are now presenting a brief summary of the workshop sessions based on the teachers' reports, the students' weekly summaries of the workshop and the students' individual summaries at the end of the term.

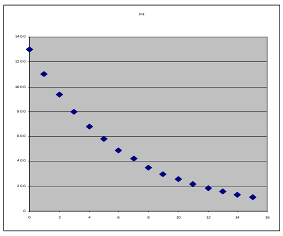
Session 1: Considering the initial question and first exploration of data

The first session is dedicated to present the generative question and the data. Each group is assigned two products from the list. During some time, the students can explore the question and propose a first forecast for the next three-month period. Most of the groups decided to introduce the data in an Excel sheet so as to represent them graphically. Most groups were able to associate the graphic representation with some of the families of functions previously studied. Some of the graphs obtained were:









Depending on the product considered, different types of functions can be associated with the graphic. The case of product 1 is different because the form of the data clearly suggests a cubic function. In this case, the students easily found an analytic expression $y = a(x - b)^3 + c$ fitting the data, first detecting the inflexion point $(b;c)$ and then testing different values for parameter a. At the end of the session, the teacher asked some of the groups to present their procedure used and results to the whole group. A structure for the Excel sheet was agreed upon and the teams were asked to bring in a possible model with its corresponding forecasts for the next session.

Session 2: Finding different models and comparing them

Each group presented the analytic expression obtained for the products assigned. As each product was assigned to different groups, different possible models appeared for the same set of data. Hence the problem of deciding which forecast was "better" quickly appeared. As it was impossible to decide on at first sight, the teacher introduced a possible criterion to 'measure how different each model was from the data'. It consists in computing the difference (in absolute value) between the values of the function and the data of the product. A new column was added to the Excel sheet (with) which, at the end, mentioned the arithmetic average of the differences. It was called the 'average error'.

Then the session work consisted in finding, for each product and within a given family of functions, the model that gives the minimum average error. The first procedure was to modify the parameters of each function to find the best model by trial and error. In the middle of the session, the teacher introduced the Excel tool 'SOLVER' that gives the parameter combination that minimizes the average error, when initial values are close to the solution. The Solver function allows finding the best approximation to data when models are considered within the same family of functions, but it is not an effective tool to decide between two models belonging to different families (a parabola and an exponential function, for instance). Besides given two sales forecasts done with functions of a different type, the fact that one of them gave a lower average error than the other, did not always seem a good criterion to determine that it was a better forecast (it is not always so clear graphically, for example). The session concluded by asking the students to bring in 'the best model' for each product and the corresponding forecast. In a sense, the first question of the initial problem was almost answered.

Session 3: Study of the average term variations

The session started by sharing the expressions provided by each group. The problem of finding a criterion to select the best model was raised in the case of different models for the same product with a similar average error. At this moment, in one of the four groups, the teacher took advantage of the work done by a team that initially, during the first session, used the term variation of the sales. They found out the rate of the previous terms' variation and then took an average to do the forecast. This idea was introduced to the rest of the teams and also to the other class groups.

Therefore, besides the data of term sales and its possible models, appear a new set of data, the term variations of the sales, which can be modelled in turn. The students were thus asked to proceed with this new data in the same way they did before: doing the graphic representation, deciding which family of functions seems to correspond to the visual tendency, finding the concrete function that gives the lower average error.

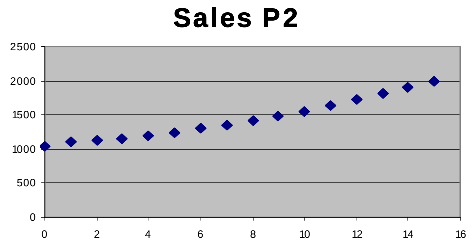
In the case of product 1 (cubic function), the new data appeared as having a quadratic tendency. In the case of products given by a quadratic function, the term variations seemed to correspond to a straight line, in the case of a rational or an exponential function, to another rational or exponential function respectively.

Sessions 4 & 5: Comparing the model of the variations to the variation of the model

When the different groups presented their models for the sales forecast and for the sales variation forecast, the teacher asked for a possible relation between the two models corresponding to the same product. In the case of the products with only one 'good model' (such as product 1 with a 'cubic tendency') the conclusion was quite complicated. With those products accepting more than one model, the variation study led to a better conclusion: the graphic that best fitted the term variations of sales was similar to the graphic of the derivative of the function that best modelled the product.

For example, if we consider product 2, we find:

Term	t	Sales
March-03	0	1050
June-03	1	1100
Sept- 03	2	1120
Dec-03	3	1160
………	…	…



The graphic representation shows a tendency that can be modelled by a linear, a quadratic or an exponential function. The corresponding average errors are:

a =	60,92
b =	973,03

t	P2	*y* =*ax*+*b*	ERROR
0	1050	973,03	76,98
1	1100	1033,95	66,05
...	...	...	...
		AVERAGE	**40,625**

a =	326,96
b =	1,10
c =	732,97

t	P2	$y = ab^x + c$	ERROR
0	1050	1059,92	9,92
1	1100	1091,30	8,70
...	...	...	...
		AVERAGE	**7,16**

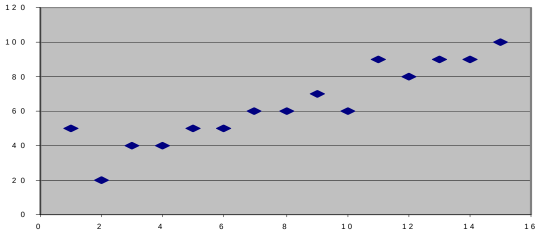
a =	2,46
b =	-5,18
c =	995,01

t	P2	$y = a(x - b)^2 + c$	ERROR
0	1050	1061,00	11,00
1	1100	1088,95	11,05
...	...	...	...
		AVERAGE	**3,63**

The study of the average error rules the linear model out, but does not provide a good criterion to exclude the exponential function or the parabola. If we consider the term variation of sales and model the new data, we obtain the following:

Time	P2	Term Var.
0	105 0	
1	110 0	50

Sales term average



a =	5,0000
b =	24,9999

t	P2	T.V.	***y* = *ax* + *b***	Error Abs
0	1050			
1	1100	50	30,00	20
2	1120	20	35,00	15
3	1160	40	40,00	0
4	1200	40	45,00	5
...	...	...	...	...
			AVERAGE	5,67

a =	-0,9915
b =	-1,2772
c =	61,0730

t	P2	T.V.	$y = ab^x + c$	Error Abs
0	1050			
1	1100	50	62,34	12,34
2	1120	20	59,46	39,46
3	1160	40	63,14	23,14
4	1200	40	58,43	18,43
...	...	...	...	...
			AVERAGE	16,93

Looking at the two corresponding term variation models, it clearly appears that the linear model has a lower average error than the exponential one. To summarize, we have found two models that fit the initial data in a similar way. Their analytic expressions, using the Excel tool 'Solver', are:

OPTION 1: $y = 326{,}96\ (1{,}09)^x + 732{,}96$ → *average error*: 7,16

OPTION 2: $y = 2{,}46\ (x + 5{,}18)^2 + 995{,}01$ → *average error*: 3,63

The lower error corresponds to the parabola, but both are similar (comparing to other considered possible models). When considering the term average of the sales, the model that fits better is: $y = 5x + 25$. Finally, if we take the first model expression $y = 2{,}46\ (x + 5{,}18)^2 + 995{,}01$ and derivate it, we get an expression very similar to the model found: $y' = 2{,}46 \cdot 2 \cdot (x + 5{,}18) = 5{,}2x + 26{,}936 \approx 5x + 25$

Therefore, we have a new criterion to decide between two models: studying both the tendency of the sales and of their term variation, and choosing as 'best model' *the* *function that has a derivative that fits the model of the term variation*. At this moment, further work on the mathematical model can follow, looking at the derivative as a model of the term variation $\Delta f(x) = f(x) - f(x - 1)$. The use of a symbolic calculator was an important tool for this final step of the modelling process, which was left to the students as a complementary theoretical analysis of the whole work done in the workshop. After these five sessions, students were able to use all the information to present a forecast for the sales and report a complete answer to the initial question.

3. THE 'MATHEMATISING STEP': CONTRASTING MODELS TO DATA

3.1. First part of the workshop: the problem of choosing the best model

The process of mathematising or assigning an appropriate mathematical model to a given system can be done in a simple way by directly choosing a previously available model given by an external source (a 'media'). However, the productivity of the model, that is, the fact that it produces new knowledge about the system, requires a certain 'fit' or 'adaptation' to the system. This process is rarely done once and for all. It requires a forth and back movement between the model and the system, in a sort of questions-answers or trial-error dynamics. We will now see the details of this process in the concrete modelling process of the workshop presented below.

In the first part of workshop, the aim is to look for a function that accurately reproduces the sales dynamic. The first decision to take is to fix the family function that seems to reproduce the observed dynamic in the data. The students' first gesture was to represent the data in a calculus sheet and determine a priori which type of function would be chosen[2]. In terms of the 'media and milieu dialectics', we can consider that the Excel graphic works as a *milieu*: when representing the chosen function, it allows to visually contrast the 'proximity' between the model and the data.

The problem about how to construct a criterion to determine the best fit is the crucial question that drives the study process. Except in one or two cases, the only visual comparison between different sales models becomes an early limited *milieu*. The necessity of establishing a 'measure of the fit' comes up, and enriches the initial *milieu* given by the numeric data series and its graphic representation. The option chosen –a new message (*media*) given by teacher– is to calculate the average of the differences (in absolute value) between the data and the values of the considered function. The incorporation of the Solver function –that works as a black box for the students– provides another *milieu* that makes the search of the function that minimizes the error more dynamic. However, this new enriched milieu can also show its limitations when the errors between different 'competitive' models are similar.

[2] The fact that students work with a small group of a pre-established family of functions does not have to be considered as a didactic limitation. It reproduces the usual situation of the genuine modelling work.

3.2. Second part: the model of the variations and the variations of the model

In the case of having different models with similar errors, the *milieu* made up of numeric values and the graphics of both sales and models is newly enriched by the introduction of a new variable: the sales variation. A new modelling process starts, similar to the previous one. The derivative function, as an approximation of the variation, soon becomes a new element of the *milieu* brought by the teacher acting as a *media*. It will contribute as a new criterion of validation: if a model fits the sales, the derivative of the model should be a good fit of the sales variation. For example, if sales seem to follow a parabolic growth, it is expected that the sales variation will follow a straight line growth. In this case, the *milieu* is all the work done during the first part of the workshop, that is, the construction of different models to each data series.

The teacher is who introduces the relation between the term variations and the derivative of the pre-established model (*media*). Besides, as students had a symbolic calculator that allowed them to easily calculate the algebraic expression of the average value $f(x + 1) - f(x)$ of any function, it was also possible to compare the derivative value of the model with this average value and confirm the approximation. It is important to underline that the increase of the '*milieu*'s complexity' made the development of this second part of the workshop more difficult, the 'system' that was to be modelled being less known and 'unstable' for the students. However, the work done represents an exemplary case of the functionality of the derivative as a simple way to calculate the average variation of a function between two points.

4. CONCLUSIONS

Using the modelling cycle proposed by Blum & Leiβ (2006), the whole process can be described in the following way. The problem of forecasting sales given a time-series of data constitutes the initial extra-mathematical situation, that we will call the 'system' (as opposed to the 'model'). At this stage, the system considered was a 'real one' (extra-mathematical). The first step of the modelling process consists in representing the data graphically to make a first hypothesis about the tendency of the time series. This first graphical model helps to decide on the type of functional model that best fits the data, giving rise to a mathematising process aimed to decide on the parameters of the chosen concrete function by a trial and error procedure using Excel, going forth and back from the model to the system. A new question arises when different types of functions are used to fit the data and one has to decide which model is best. The search for a criterion needs to consider a new 'real system' formed by the data and the possible models, with the problematic question of how to determine the 'best fit', that is, how to mathematically model the 'fitness' of a model. This new system is in turn mathematised by the average error of the fit. Again, the insufficiencies of this new model lead to the consideration of a new enriched 'system': the one formed by the original data and the term variation of the sales. A possible criterion is set up by considering the double modelling of the sales and the term variation of the sales. Finally, considering the derivative as a model of the term variation constitutes the last mathematisation step that leads to a final conclusion for the forecast problem.

It is important to note that, in this entire process, the successive ‘systems’ that are modelled are more and more mathematised, and that the successive ‘models’ constructed progressively integrate the previous systems, creating new problems and, thus, generating the need to go on with the modelling process. We have interpreted these successive mathematising processes using the ‘milieu and media dialectics’ introduced by Chevallard (2004), which has helped us provide a detailed analysis of the mathematising step of the modelling process, showing how being a ‘system’ to be modelled or a ‘model’ of the system is more related to the *function* assigned to a given object during the modelling process than to its very ‘nature’ (it being mathematical or extra-mathematical). The example here described shows how the development of a modelling activity, even if initiated with an extra-mathematical situation, leads to consider, not only a sequence of new models, but also new and enriched systems more and more mathematised. Hence, extra-mathematical and intra-mathematical modellings appear as strongly intertwined.

REFERENCES

Blum, W.& Leiß, D. (2006) “Filling up“ – The problem of independence-preserving teacher interventions in lessons with demanding modelling tasks. In Bosch, M. (Ed.) *Proceedings of the 4th Conference of the European Society for Research in Mathematics Education (CERME 4)*. Barcelona: FUNDEMI IQS.

Brousseau, G. (1997) *Theory of didactical situations in mathematics: Didactique des mathématiques* 1970-1990 (N. Balacheff, M. Cooper, R. Sutherland and V. Warfield, Eds. and Trans.). Dordrecht, The Netherlands: Kluwer.

Chevallard, Y. (2004). Vers une didactique de la codisciplinarité. Notes sur une nouvelle épistémologie scolaire. *Journées de didactique comparée* (Lyon, mai 2004). (http://yves.chevallard.free.fr)

Chevallard, Y. (2006), Steps towards a new epistemology in mathematics education. In Bosch, M. (ed.) *Proceedings of the 4th Conference of the European Society for Research in Mathematics Education (CERME 4).* Sant Feliu de Guíxols, Spain. (pp. 21-30).